Beszédadatbázis irodai számítógép-felhasználói környezetben

Vicsi Klára [1], Kocsor András [2], Teleki Csaba[1], Tóth László[2]

[1] BME Távközlési és Médiainformatikai Tanszék, Beszédakusztikai Kutatólaboratórium, Sztoczek u. 2., 1111 Budapest, Magyarország {vicsi, teleki}@tmit.bme.hu
ttp://alpha.tmit.bme.hu/speech
[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport, Aradi vértanuk tere 1., 6720 Szeged, Magyarország
{kocsor, tothl}@inf.u-szeged.hu

Kivonat: Az előadásban bemutatjuk az új olvasott szöveget tartalmazó Magyar Referencia Beszédadatbázist, amelyet általános felhasználói környezetben, irodákban, laboratóriumokban, lakásokban rögzítettünk, összefoglaljuk az adatbázis létrehozásának akusztikai, nyelvi feldolgozási módszereit, ismertetjük az adatbázis pontos jellemzőit. Az adatbázis 332 beszélő olvasott szövegét tartalmazó hanganyag, beszélőnként 12 mondat és 12 szó speciális fonetikai elvárásoknak megfelelően összeállítva került bemondásra. A felvételek többféle mikrofonnal, hangkártyával és személyi számítógéppel készültek.

1 Bevezetés

Célunk egy általános felhasználású, irodai, otthoni környezetben olvasott folyamatos szöveget tartalmazó, beszédadatbázis létrehozása és akusztikai, valamint nyelvi feldolgozása, amely alkalmas PC-s beszédfelismerők betanítására, tesztelésére. A létrehozott beszéd adatbázist *Magyar Referencia Beszédadatbázis*nak *(MRBA)* neveztük el. A BME TMIT Beszédakusztikai Laboratóriuma a szegedi SZTE Informatikai Tanszékcsoporttal együttműködve hozza létre ezt az adatbázist. A BME TMIT Beszédakusztikai Laboratóriuma végezte el a beszédadatbázis szöveganyagának megtervezését, az előfeldolgozást és annotálást, az automatikus betű – fonéma átírást, a szegedi SZTE Informatikai Tanszékcsoport a kézi szegmentálást végezte, a felvételek készítése megosztva történt.

1.1 A beszédadatbázis szöveganyagának megtervezése

A létrehozandó adatbázis számos különböző típusú beszédfelismerő betanítására és tesztelésére kell hogy lehetőséget adjon. Mivel a felismerés alapja ma már szavaknál kisebb egység, fonéma, difon, trifon, stb., olyan folyamatos szöveg összeállítására van szükség, ahol ezek az elemek elegendően sokszor fordulnak elő. A beszédfelismerési célokra rögzítendő beszédnek a lehető legjobban kell fednie a beszélt magyar nyelv

sajátosságait. Mivel a felvett szöveg minden másodperce sok munkát von maga után, a szöveganyagnak minél rövidebbnek is kell lennie. Ezért alapos statisztikai vizsgálatokra van szükség a fonémák, difonok, trifonok szintjén egyaránt. Ennek alapján kell összeállítani olyan folyamatos szöveget, ahol a leggyakoribb bi- és trifonok megfelelő mennyiségben állnak rendelkezésre.

A szöveget fonémák sorozatára alakítottuk egy speciális algoritmus segítségével. Fonotipikus fonetikai átírást alkalmaztunk (Vicsi 2001, Fourcin, A.J. and Dolmazon, J-M.1991), vagyis a karakterek átírását a nyelv fonetikai szabályainak alapján végeztuk el úgy, hogy a szövegkörnyezetet is figyelembe vettük (pl. koartikuláció, hasonulás).

Az ortografikus karakterek fonetikai átírásához SAMPA fonetikai jelölést használtunk (Vicsi, 2000). A fonémaátírás után következett a statisztikai vizsgálat. A fonéma, bifon és trifon megoszlást vizsgáltuk az eredeti 1,6 MB méretű szövegadatbázison.

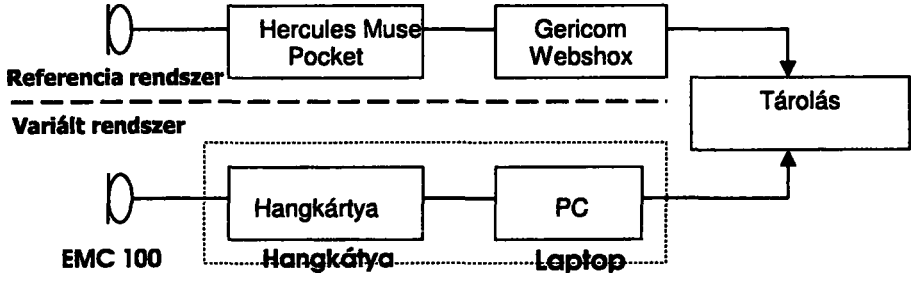
Egy bemondó 12 különböző mondatot és 12 különböző, a mondatoktól független, szót olvas fel. A teljes szöveganyag 166 bemondásra készült el azért, hogy szöveganyag 2-szer kerülhessen ismétlésre a 332 bemondás során. Így a teljes szöveganyag 12 x 166 azaz összesen 1992 db különböző mondatból és 12 x 166, azaz 1992 db különböző, a mondatoktól független szóból áll.

A mondatok kiválasztása a következő szempontok szerint történt: a fonémastatisztika alapján az 1%-nál gyakoribb fonémákat legalább egy példányban tartalmaznia kellett minden csoport valamelyik mondatának.

A teljes szöveg és a válogatott szöveg bifon statisztikáinak alapján a leggyakoribb 98.8% bifon mindegyike szerepel mindkét anyagban, és csak itt kezdenek el hiányozni a válogatott szövegből a bifonok.

2 A beszédadatbázis rögzítése és feldolgozása

A beszédadatbázis felvételeit különböző helyszíneken: zajos, kevésbé zajos iroda



1. ábra Az MRBA adatbázis felvételi elrendezése

helyiségekben, laborokban, otthonokban vettük fel. A felvételeknél szinkronban két különböző rendszerrel dolgoztunk. Az egyik az ún. *referenciarendszer*, ahol jó minőségű, közelbeszélő kondenzátor mikrofont (Monacor EMC 100), jó technikai paraméterekkel rendelkező hangkártyát (Hercules Muse Pocket USB 5.1), valamint egy adott Gericom Webshox típusú laptopot használtunk. A másik rendszernél az ún.

variált rendszer, különböző, jobb, kevésbé jó mikrofonokat, hangkártyákat, PC-ket használtunk, a lehető legnagyobb variáltsággal. A felvételi elrendezést az 1. ábra mutatja. Felvételeknél a bemondók, a PC-k, a hangkártyák, a mikrofonok és a környezet minél nagyobb variáltságára törekedtünk.

2.1 A beszélők demográfiai adatai

Régiók és dialektusok: A felvételeket Magyarország négy különböző tájegységében lévő városban rögzítettük: Budapesten, Szegeden, Győrben és Miskolcon. A felvételek során a beszélők születési helyét és jelenlegi lakhelyét jegyeztük fel.
Táblázat 3. A beszélők életkor és neme szerinti megoszlása

Korcsoport	Férfi beszélők	Női beszélők
16 év alatt	0,9 %	3,3 %
16 – 30 év	46,1 %	27,7 %
31 – 45 év	5,7 %	6 %
46 – 60 év	3,9 %	5,1 %
60 év felett	0,9 %	0,4 %
Összesen	*57,5 %*	*42,5 %*

2.2 Előfeldolgozás és annotálás

Az előfeldolgozás a felvételek lehallgatásából, ellenőrzéséből és esetleg feldarabolásából, esetünkben a két csatorna szinkronizálásából áll.
Az annotálás annyit jelent, hogy minden hangfájl mellé egy címkefájlt készítünk, amely különféle információkat tartalmaz a hangfájl paramétereivel és tartalmával kapcsolatban: az elhangzott szöveg ortografikus lejegyzését, hibás kiejtést, nem érthető szavakat, szótöredékeket, a beszélő nem beszédből származó hangjait, környezeti zajokat, stb. (Wells, J. 2001). Az alábbi ábrában egy ilyen cimkefájlt prezentálunk.

2.3 Fonotipikus automatikus betű-fonéma átírás, szóhatár megadás

A bemondott szöveg betűit átírtuk úgy, hogy figyelembe vettük a magyar beszéd fonetikai szabályait a szövegkörnyezet függvényében. Ilyenek például az alkalmazkodási folyamatok (hasonulás, asszimiláció, stb.).
A folyamatos beszédre tipikusan jellemző, hogy a szavak között nincs szünet. A fizikai paraméterek folyamatosan változnak a szóhatárokon. Ezért a további feldolgozás megkönnyítése érdekében bejelöltük a szavak határait is.

2.4 Kézi szegmentálás

Az adatbázis annotálása után következő feladat annak fonetikai szintű szegmentálása és címkézése, valamint a szóhatárok és a frázishatárok bejelölése. A feladat „audiovizuális fonetikai átírás" a 1997-ban elkészült BABEL nemzetközi project leírása

alapján (Vicsi K., Vig A.). Az átírást a szöveg hallgatása, és az időfüggvény és/vagy a színkép elemzése alapján hajtottuk végre.
Az aktuális kimondásnál kimaradt hangokat megjelöltük, a követő hang előtt zárójelbe tettük.
A **szószintű szegmentálás** esetén a szavak határát jelöltük. A **mondatok határait** a frázisjelző írásjelek (vessző, kettőspont, pontosvessző, gondolatjel, macskaköröm, és, pont kérdőjel, felkiáltójel) adták.

3. A Magyar Referencia Beszédadatbázis adatkészleteinek átfogó műszaki tulajdonságai

Magyar nyelvű, olvasott szövegű, személyi számítógépes környezetben felvett adatbázis, 16 bites, 16 kHz-es mintavételezéssel.

- *332 beszélő közvetlenül a számítógépbe rögzített hanganyaga;*
- *Beszélőnként 12 mondat és 12 szó*
- *A felvételek többféle mikrofonnal, hangkártyával, PC-kel készültek*
- *Környezet változó zajosságú irodahelyiség, laboratórium, otthoni környezet;*
- *Az adatbázis teljes anyaga annotált, az adatbázis harmada (100 beszélő) kézileg szegmentált és címkézett.*

4. Köszönetnyilvánítás

A kutatás az OTKA T 046487 ELE és az IKTA 00056 pályázatok keretén belül készül.

5. Irodalom

Vicsi, K. Beszédadatbázisok a gépi beszédfelismerés segítésére, Híradástechnika, Vol. 2001/1, Budapest, pp. 5-13, 2001.

Vicsi, K. Vig, A.: Az első magyar nyelvű beszédadatbázis, Beszédkutatás'98, Tanulmányok az elméleti és alkalmazott fonetika köréből. MTA Nyelvtudományi Intézet, Budapest, pp. 163-178, 1998.

Wells, J. at all.: Standard Computer-Compatible Transcription. Esprit Project 2589 (SAM), Doc. no. SAM-UCL-037. London: Phonetics and Linguistics Dept., UCL (1992).

Fourcin, A.J. and Dolmazon, J-M. „Speech knowledge, standards and assessment", Proceedings of XII International Congress of Phonetic Sciences, Aix-en-Provence, Vol. 5, 430-433 (1991).